

Linguistically Regularized LSTMs for Sentiment Classification

Qiao Qian, Minlie Huang, Xiaoyan Zhu

State Key Lab. of Intelligent Technology and Systems, National Lab. for Information Science and Technology
Dept. of Computer Science and Technology, Tsinghua University, Beijing 100084, PR China
qianqiaodecember29@126.com, aihuang@tsinghua.edu.cn, zxy-dcs@tsinghua.edu.cn

Abstract

Sentiment understanding has been a long-term goal of AI in the past decades. This paper deals with sentence-level sentiment classification. Though a variety of neural network models have been proposed very recently, however, previous models either depend on expensive phrase-level annotation, whose performance drops substantially when trained with only sentence-level annotation; or do not fully employ linguistic resources (e.g., sentiment lexicons, negation words, intensity words), thus not being able to produce linguistically coherent representations. In this paper, we propose simple models trained with sentence-level annotation, but also attempt to generating linguistically coherent representations by employing regularizers that model the linguistic role of sentiment lexicons, negation words, and intensity words. Results show that our models are effective to capture the sentiment shifting effect of sentiment, negation, and intensity words, while still obtain competitive results without sacrificing the models' simplicity.

Introduction

Understanding sentiment has always been one of the goals of AI in the decades. As a small step toward sentiment understanding, sentiment classification aims to classify sentiment to sentiment classes such as *positive or negative*, or more fine-grained classes such as *very positive*, *positive*, *neutral*, *etc.* There has been a variety of approaches for this purpose such as lexicon-based classification (Turney 2002), and early machine learning based methods (Pang, Lee, and Vaithyanathan 2002; Pang and Lee 2005), and recently neural network models such as convolutional neural network (CNN) (Kim 2014; Kalchbrenner, Grefenstette, and Blunsom 2014), recursive autoencoders (Socher et al. 2011; Socher et al. 2013), Long Short-Term Memory (LSTM) (Mikolov 2012; Chung et al. 2014; Tai, Socher, and Manning 2015; Zhu, Sobhani, and Guo 2015), and many more.

In spite of the great success of these neural models, there are some defects in previous studies. First, tree-structured models such as recursive autoencoders and Tree-LSTM (Tai, Socher, and Manning 2015; Zhu, Sobhani, and Guo 2015), depend on parsing tree structures and expensive phrase-level

annotation, whose performance drops substantially when only trained with sentence-level annotation. Second, sequence models such as CNN and recurrent network are not easy to produce competitive results as reported in the literature (Tai, Socher, and Manning 2015). Third, linguistic knowledge has not been fully employed in neural models, though (Qian et al. 2015) shows that part-of-speech tags can be quite effective for sentence-level classification.

The goal of this research is to developing simple sequence models but also attempts to fully employing linguistic resources to benefit sentiment classification. Firstly, we attempt to develop simple models that do not depend on parsing trees and avoid phrase-level annotation which is too expensive in real-world applications. Secondly, in order to obtain competitive performance, simple models can benefit from linguistic resources. Three types of resources can be addressed: sentiment lexicon, negation words, and intensity words. Sentiment lexicon offers the prior polarity of a word which can be useful in determining the sentiment polarity of longer texts such as phrases and sentences. Negation words are typical sentiment shifters (Zhu et al. 2014), which constantly shift sentiment expression. Intensify words such as *very*, *and extremely* change the valence degree of sentiment, which is important for fine-grained sentiment classification.

In order to model the linguistic role of sentiment, negation, and intensity words, and thus to generate linguistically coherent representations, our central idea is to regularize the difference between the predicted sentiment distribution of the current position, and that of the previous or next positions. For instance, if the current position is a negation word *not*, the current predicted distribution should be close to the transformed distribution of the next predicted distribution parameterized by the negation transformation matrix. To summarize, our contributions lie in three folds:

- We propose several regularizers to model the linguistic role of sentiment, negation, and intensity words in sentiment classification. Experiments show that the regularizers are quite effective.
- Due to the complexity of sentiment shifting effect of negation and intensify words, we design word-specific transformation matrices to respect the role of each word.
- Unlike previous models depend on parsing structures and expensive phrase-level annotation, our models are simple

and efficient but also obtain competitive performance.

Related Work

Neural Networks for Sentiment Classification Recently, there are many neural networks proposed for sentiment classification. The most pioneering (perhaps) model may be the recursive autoencoder neural network which builds the representation of a sentence from subphrases recursively (Socher et al. 2011; Socher et al. 2013; Dong et al. 2014; Qian et al. 2015). Such recursive models usually depend on a tree structure of input text, and in order to obtain competitive results, usually require heavy annotation on each subphrase. Sequence models do not depend on a particular tree structure, for instance, convolutional neural network (CNN) is another type of widely used models for sentiment classification (Kim 2014; Kalchbrenner, Grefenstette, and Blunsom 2014). As used similarly in image processing, CNN defines convolution operations on a sequence of a text. Long short-term memory models are also common for learning sentence-level representation due to its capability of modeling the prefix or suffix context (Hochreiter and Schmidhuber 1997). LSTM can be commonly applied to sequential data but also tree-structured parsing trees (Zhu, Sobhani, and Guo 2015; Tai, Socher, and Manning 2015). A complete search for optimal structures of recurrent networks and LSTMs can be seen in (Chung et al. 2014).

Applying Linguistic Knowledge for Sentiment Classification Linguistic knowledge and sentiment resources are very helpful for sentiment analysis such as sentiment lexicons, negation words (*not*, *never*, *neither*, *etc.*), and intensity words (*very*, *extremely*, *etc.*).

Sentiment lexicon, such as Hu and Liu Lexicon (Hu and Liu 2004) and MPQA lexicon (Wilson, Wiebe, and Hoffmann 2005), is widely used for sentiment classification (Pang and Lee 2008).

Negation words play a critical role in modifying sentiment of textual expressions. Some early negation models are designed to reverse the sign of sentiment value of the modified text (Polanyi and Zaenen 2006). Since each individual negation word can affect sentiment words in different ways, the shifting hypothesis, is proposed, assuming that negators change the sentiment values by a constant amount (Taboada et al. 2011). The modified word is also an important factor on the negation effect, for instance, negation words turn positive to negative and turn negative to not negative. (Zhu et al. 2014) incorporate negation words as feature into neural network. (Kiritchenko and Mohammad 2016) incorporate negation words and other linguistic knowledge into a SVM classifier for composing opposing polarities.

The intensity words can change the valence degree (i.e., sentiment intensity) of the content word. Sentiment intensity of a phrase indicates the strength of associated sentiment, which is quite important for fine-grained sentiment classification or rating. (Taboada et al. 2011) directly reverse the polarity of modified words or change the sentiment strength by a fixed value. (Wei, Wu, and Lin 2011) predict the valence value for content words using a linear regression model. (Malandrakis et al. 2013) introduce a kernel function

to combine semantic information for predicting sentiment score. In the SemEval-2016 task 7 subtask A, (Wang, Zhang, and Lan 2016) propose a learning-to-rank model with a pairwise strategy to predict sentiment intensity scores.

Long Short-term Memory Network

Long Short-Term Memory (LSTM)

To deal with the notorious issues of gradient explosion and vanishing in recurrent neural network (Hochreiter and Schmidhuber 1997), Long Short-term Memory network is proposed by incorporating an additional memory cell $c_t \in R^d$ at each time step. The hidden states h_t and memory cell c_t is a function of their previous c_{t-1} and h_{t-1} and input vector x_t , formally defined as follows:

$$c_t, h_t = g^{(LSTM)}(c_{t-1}, h_{t-1}, x_t) \quad (1)$$

The hidden state $h_t \in R^d$ denotes the representation of position t while also encoding the preceding context of the position. For more details about LSTM, we refer readers to (Chung et al. 2014).

Bidirectional LSTM

In LSTM, the representation of each position (h_t) only encodes the prefix context in a forward direction while the backward context is not respected. Bidirectional LSTM (Graves, Jaitly, and Mohamed 2013) exploited two parallel passes (forward and backward) and concatenated representations of the two LSTMs as the representation of each position. The forward and backward LSTMs are respectively formulated as follows:

$$\vec{c}_t, \vec{h}_t = g^{(LSTM)}(\vec{c}_{t-1}, \vec{h}_{t-1}, x_t) \quad (2)$$

$$\overleftarrow{c}_t, \overleftarrow{h}_t = g^{(LSTM)}(\overleftarrow{c}_{t+1}, \overleftarrow{h}_{t+1}, x_t) \quad (3)$$

where $g^{(LSTM)}$ is the same as that in Eq 1. Particularly, parameters in the two LSTMs are shared. The representation of the entire sentence is $[\vec{h}_n, \overleftarrow{h}_1]$, where n is the length of the sentence. At each position t , the joint representation $h_t = [\vec{h}_t, \overleftarrow{h}_t]$, which is the concatenation of hidden states of the forward LSTM and backward LSTM. In this way, the forward and backward contexts can be considered simultaneously.

Linguistically Regularized LSTM

The central idea of the paper is to output linguistically coherent predictions in sentiment classification by regularizing the outputs at adjacent positions of a sentence. For example, in sentence “this movie is interesting”, the predicted sentiment distributions at “this*¹”, “this movie*”, and “this movie is*” should almost be the same, while the predicted sentiment distribution at “this movie is very interesting*” should be quite different from the preceding positions since a sentiment word (“interesting”) is seen.

More formally, the predicted sentiment distribution (p_t , based on h_t , see Eq. 4) at position t should be linguistically

¹The asterisk denotes the current position.

regularized with respect to that of the preceding $(t - 1)$ or following $(t + 1)$ positions. We propose a generic regularizer and three special regularizers based on the following linguistic observations:

- **Non-Sentiment Regularizer:** if the two adjacent positions are all non-opinion words, the sentiment distributions of the two positions should be close to each other.
- **Sentiment Regularizer:** if the word is a sentiment word found in a lexicon, the sentiment distribution of the current position should be significantly different from that of the next or previous positions.
- **Negation Regularizer:** Negation words such as “not” and “never” are critical sentiment shifter (Kennedy and Inkpen 2006): usually shifts sentiment polarity from the positive side to the negative one, but sometimes highly depends on the negation word and the words they modify. The negation regularizer models this linguistic phenomena.
- **Intensity Regularizer:** Intensity words such as “very” and “extremely” change the valence degree of a sentiment expression: for instance, from *positive* to *very positive*. Modeling this effect is quite important for fine-grained sentiment classification, and the intensity regularizer is designed to formulate this effect.

In order to enforce the model to produce coherent predictions, we propose a new loss function as follows to incorporate these regularizers:

$$E(\theta) = - \sum_i y^i \log p^i + \alpha \sum_i \sum_t L_t^i + \beta ||\theta||^2 \quad (4)$$

where y^i is the gold distribution, p^i is the predicted distribution output from a softmax layer taking the sentence representation as input, L_t^i is one of the above regularizers or combination of these regularizers, α is the weight for the regularization term, and i, t is the index of sentence and position respectively.

Non-Sentiment Regularizer (NSR)

This regularizer constrains that the sentiment distributions of adjacent positions should not vary much if the additional input word x_t is not a sentiment word, formally as follows:

$$L_t^{(NSR)} = \max(0, D_{KL}(p_t, p_{t-1}) - M) \quad (5)$$

where M is a hyperparameter for margin, p_t is the predicted distribution at position t whose representation is h_t , and $D_{KL}(p, q)$ is a symmetric KL divergence defined as follows:

$$D_{KL} = \frac{1}{2} \sum_{l=1}^C p(l) \log q(l) + q(l) \log p(l)$$

where p, q are distributions over sentiment labels.

Sentiment Regularizer (SR)

The sentiment regularizer constrains that the sentiment distributions of adjacent positions should drift accordingly if the input word is a sentiment word. Let’s revisit the example

“this movie is interesting” again. At position $t = 4$ we see a positive word “interesting” so the predicted distribution at this position would be more positive than that at position $t = 3$. This is the issue of *sentiment drift*.

In order to address the sentiment drift issue, we propose a polarity shifting distribution $s_c \in R^C$ for each sentiment class defined in a lexicon. For instance, a sentiment lexicon may have class labels like *strong positive*, *weakly positive*, *weakly negative*, and *strong negative*, and for each class, there is a shifting distribution which will be learned by the model. The sentiment regularizer states that if the current word is a sentiment word, the sentiment distribution drift should be observed in comparison to the previous position, as formulated as follows:

$$p_{t-1}^{(SR)} = p_{t-1} + s_{c(x_t)} \quad (6)$$

$$L_t^{(SR)} = \max(0, D_{KL}(p_t, p_{t-1}^{(SR)}) - M) \quad (7)$$

where $p_{t-1}^{(SR)}$ is the drifted sentiment distribution after considering the shifting sentiment distribution corresponding to the word at position t , $c(x_t)$ is the prior sentiment class of word x_t , and $s_c \in \theta$ is a parameter to be optimized but could also be set fixed with prior knowledge. Note that in this way all words of the same sentiment class share the same drifting distribution, but in a refined setting, we can learn a shifting distribution for each sentiment word if large-scale datasets are available.

Negation Regularizer (NR)

The negation regularizer approaches how negation words shift the sentiment distribution of its modifiers. When the input word x_t is a negation word, the sentiment distribution should be shifted accordingly. However, the negation role is more complex than that by sentiment words, for example, the word “not” in “not good” and “not bad” have different roles in polarity shifting. The former changes the polarity to *negative*, while the latter changes to *neutral* instead of *positive*.

To respect such complex negation effects, we propose a transformation matrix $T_m \in R^{C \times C}$ for each negation word m , and the matrix will be learned by the model. The regularizer assumes that if the current position is a negation word, the sentiment distribution of the current position should be close to that of the next or previous position with the transformation.

$$p_{t-1}^{(NR)} = \text{softmax}(T_{x_j} \times p_{t-1}) \quad (8)$$

$$p_{t+1}^{(NR)} = \text{softmax}(T_{x_j} \times p_{t+1}) \quad (9)$$

$$L_t^{(NR)} = \min \begin{cases} \max(0, D_{KL}(p_t, p_{t-1}^{(NR)}) - M) \\ \max(0, D_{KL}(p_t, p_{t+1}^{(NR)}) - M) \end{cases} \quad (10)$$

where $p_{t-1}^{(NR)}$ and $p_{t+1}^{(NR)}$ is the sentiment distribution after transformation, $T_{x_j} \in \theta$ is the transformation matrix for a negation word x_j , a parameter to be learned during training. In total, we train m transformation matrixes for m negation words.

Intensity Regularizer (IR)

The intensify regularizer models how intensity words influence the sentiment valence of a phrase or a sentence. Intensifier can change the valence degree of the content word. Sentiment intensity of a phrase indicates the strength of associated sentiment, which is quite important for fine-grained sentiment classification or rating.

The formulation of the intensity effect is quite the same as that in the negation regularizer, but with different parameters of course. For each intensity word, there is a transform matrix to favor the different roles of various intensifiers on sentiment shift. For brevity, we will not repeat the formulas here.

Applying Linguistic Regularizers to Bidirectional LSTM

To make our model simple and elegant in a mathematical form, we do not consider the modification scope of negation and intensity word, which is a quite challenging problem in the NLP community. However, we can alleviate the problem by leveraging bidirectional LSTM.

For a single LSTM, we employ a backward LSTM from the end to the beginning of a sentence. This is because, at most times, the modified words of negation and intensity words are usually at the right side of the modifiers. But sometimes, the modified words are at the left side of negation and intensity words. To better address this issue, we employ bidirectional LSTM and let the model determine which side should be chosen.

More formally, in Bi-LSTM, we compute a transformed sentiment distribution on \vec{p}_{t-1} of the forward LSTM and also that on \overleftarrow{p}_{t+1} of the backward LSTM, and compute the minimum distance of the distribution of the current position to the two distributions. This could be formulated as follows:

$$\vec{p}_{t-1}^{(R)} = softmax(T_{x_j} \times \vec{p}_{t-1}) \quad (11)$$

$$\overleftarrow{p}_{t+1}^{(R)} = softmax(T_{x_j} \times \overleftarrow{p}_{t+1}) \quad (12)$$

$$L_t^{(R)} = \min \begin{cases} \max(0, D_{KL}(\vec{p}_t, \vec{p}_{t-1}^{(R)}) - M) \\ \max(0, D_{KL}(\overleftarrow{p}_t, \overleftarrow{p}_{t+1}^{(R)}) - M) \end{cases} \quad (13)$$

where $\vec{p}_{t-1}^{(R)}$ and $\overleftarrow{p}_{t+1}^{(R)}$ are the sentiment distributions transformed from the previous state \vec{p}_{t-1} and next state \overleftarrow{p}_{t+1} respectively. Note that $R \in \{NR, IR\}$ indicating the formulation works for both negation and intensity regularizers.

Due to the same consideration, we redefine $L_t^{(NSR)}$ and $L_t^{(ISR)}$ with bidirectional LSTM similarly. The formulation is the same and omitted for brevity.

Discussion

Unlike previous studies on negation and intensity words, which modulate the linguistic effect of these words by some predefined rules, our models address these factors with mathematical operations, parameterized with shifting distribution vectors and transformation matrices. In the sentiment regularizer, the sentiment shifting effect is parameterized

with a *class-specific* distribution (but could also be word-specific if with more data). In the negation and intensity regularizers, the effect is parameterized with *word-specific* transformation matrices, meaning that different words have different parameters. Since the mechanism of how negation and intensity words shift sentiment expression is quite complex and highly dependent on individual words, we believe such mathematical operation will be more suitable for addressing complex linguistic roles of these words. This is a major advantage of our approach over other methods.

Experiment

Dataset and Sentiment Lexicon

Two datasets are used for evaluating the proposed models: Movie Review (MR) (Pang and Lee 2005) which has two classes as *negative*, *positive* and Stanford Sentiment Treebank (SST) (Socher et al. 2013) which has five classes. For details, we refer readers to the two papers. SST has provided phrase-level annotation on all inner nodes, but we only use the sentence-level annotation since one of our goals is to avoid expensive phrase-level annotation.

The sentiment lexicon contains two parts. The first part comes from MPQA (Wilson, Wiebe, and Hoffmann 2005), which contains 5,153 sentiment words, each with polarity rating. The second part consists of the leaf nodes of the SST dataset (i.e., all sentiment words) and there are 6,886 polar words except *neural* ones. We combine the two parts and ignore those words that have conflicting sentiment labels, and produce a lexicon of 9,750 words with 4 sentiment labels. For negation and intensity words, we collect them manually since the number is small, some of which can be seen in Table 2.

Dataset	MR	SST
# sentences in total	10,662	11,885
#sen containing sentiment word	10,446	11,211
#sen containing negation	1,644	1,832
#sen containing intensity	2,687	2,472

Table 1: The data statistics for the MR and SST datasets.

Negation word	no, not, never, neither, nothing, seldom, hardly, is not, cannot, do not
Intensity word	very, greatly, absolutely, too, completely, terribly, fairly, highly, more, really, most

Table 2: Examples of the negation and intensity words. For the complete list refer to *Supplemental Material*.

Regarding the parameters and initialization details of the models, please refer to the “Supplemental Material” due to the length limit.

Overall Comparison

We include several lines of baselines in the evaluation. The baselines are listed as follows:

- **RNN/RNTN** Recursive Neural Network over parsing trees, proposed by (Socher et al. 2011) and Recursive Tensor Neural Network (Socher et al. 2013) employs tensors to model correlations between different dimensions of child nodes’ vectors.
- **LSTM/Bi-LSTM** Long-short Term Memory (Cho et al. 2014) and the bidirectional variant as introduced previously.
- **Tree-LSTM** Tree-Structured Long Short-Term Memory (Tai, Socher, and Manning 2015) introduces memory cells and gates into tree-structured neural network.
- **CNN** Convolutional Neural Network (Kalchbrenner, Grefenstette, and Blunsom 2014) generates sentence representation by convolution and pooling operations.

Firstly, we evaluate our model on the MR dataset and the results are shown in Table 3. As can be seen, we can make the following statements:

- Both LR-LSTM and LR-Bi-LSTM outperforms their counterparts substantially (81.5% vs. 77.4% and 82.1% vs. 79.3%, resp.), demonstrating the effectiveness of the linguistic regularizers.
- LR-LSTM and LR-Bi-LSTM perform slightly better than Tree-LSTM but Tree-LSTM leverages a constituency tree structure while our model is a simple sequence model. As future work, we will apply such regularizers to tree-structured models.
- On this dataset, our model is comparable to CNN.

For fine-grained sentiment classification, we evaluate our model on the SST dataset which has five sentiment classes { *very negative*, *negative*, *neutral*, *positive*, *very positive* } so that we can evaluate the sentiment shifting effect of intensity words. The experiment result is shown in Table 3. We have the following observations:

- Similarly, linguistically regularized LSTM and Bi-LSTM are better than their counterparts. It’s worth noting that LR-Bi-LSTM (trained with just sentence-level annotation) is even comparable to Bi-LSTM trained with phrase-level annotation. That means, LR-Bi-LSTM can avoid the heavy phrase-level annotation but still obtain competitive results.
- Our models are comparable to Tree-LSTM but our models are not dependent on a parsing tree and more simple, and hence more efficient. Further, for Tree-LSTM, the model is heavily dependent on phrase-level annotation, otherwise the performance drops substantially (from 51% to 48.1%).
- On this dataset, our model is apparently better than CNN.

The Effect of Different Regularizers

In order to reveal the effect of each individual regularizer, we conduct ablation experiments. Each time, we remove a regularizer and observe how the performance varies. First of all,

Method	MR	SST	SST
		phrase-level	sentence-level
RNN	77.7	44.8	43.2
RNTN	75.9	45.7	43.4
LSTM	77.4	47.2	45.6
Bi-LSTM	79.3	49.1	46.5
Tree-LSTM	80.7	51.0	48.1
CNN	81.5	48.0	46.9
LR-Bi-LSTM	82.1	-	48.6
LR-LSTM	81.5	-	48.2

Table 3: The accuracy on Movie Review (MR) and Stanford Sentiment Treebank (SST). *Phrase-level* means the models use sentiment label for all inner nodes. And *sentence-level* means the models only use sentence-level annotation.

we conduct this experiment on the entire datasets, and then we experiment with sub-datasets that only contain negation words or intensity words.

The experiment results are shown in Table 4 where we can see that the non-sentiment regularizer and sentiment regularizer play a key role², and the negation regularizer and intensity regularizer are effective but less important than the previous two regularizers. This may be due to the fact that only 14% of sentences contains negation words in the test datasets, and 23% contains intensity words, and thus we further evaluate the models on two subsets, as shown in Table 5.

Method	MR	SST
LR-Bi-LSTM	82.1	48.6
LR-Bi-LSTM (-NSR)	80.8	46.9
LR-Bi-LSTM (-SR)	80.6	46.9
LR-Bi-LSTM (-NR)	81.2	47.6
LR-Bi-LSTM (-IR)	81.7	47.9
LR-LSTM	81.5	48.2
LR-LSTM (-NSR)	80.2	46.4
LR-LSTM (-SR)	80.2	46.6
LR-LSTM (-NR)	80.8	47.4
LR-LSTM (-IR)	81.2	47.4

Table 4: The accuracy for LR-Bi-LSTM and LR-LSTM without each regularizer. *NSR*, *SR*, *NR* and *IR* denotes *Non-sentiment Regularizer*, *Sentiment Regularizer*, *Negation Regularizer*, and *Intensity Regularizer* respectively.

The experiments on the subsets show that: 1) With linguistic regularizers, LR-Bi-LSTM outperforms Bi-LSTM remarkably on these subsets; 2) When the negation regularizer is removed from the model, the performance drops significantly on both MR and SST subsets; 3) Similar observations can be made regarding the intensity regularizer.

²Kindly note that almost all sentences contain sentiment words, see Tab. 1.

Method	Neg. Sub.		Int. Sub.	
	MR	SST	MR	SST
BiLSTM	72.0	39.8	83.2	48.8
LR-Bi-LSTM (-NR)	74.2	41.6	-	-
LR-Bi-LSTM (-IR)	-	-	85.2	50.0
LR-Bi-LSTM	78.5	44.4	87.1	53.2

Table 5: The accuracy on the negation sub-dataset (Neg. Sub.) that only contains negators and intensity sub-dataset (Int. Sub.) that only contains intensifiers.

The Effect of the Negation Regularizer

To further reveal the linguistic role of negation words, we compare the predicted sentiment distributions of a phrase pair with or without a negation word. The experimental results performed on MR are shown in Fig. 1. Each dot denotes a phrase pair (for example, $\langle \text{interesting}, \text{not interesting} \rangle$), where the x-axis denotes the positive score³ of the phrase without negators (e.g., *interesting*), and the y-axis indicates the positive score for the phrase with negators (e.g., *not interesting*). The curves in the figures show this function: $[1 - y, y] = \text{softmax}(T_{nw} * [1 - x, x])$ where $[1 - r, r]$ is a sentiment distribution on $[\text{negative}, \text{positive}]$, x is the positive score of the phrase without negators (x-axis) and y that of the phrase with negators (y-axis), and T_{nw} is the transformation matrix for the negation word nw , see Eq. 8.

We can observe the following statements:

- All the dots are distributed around the curve, and there is no dot at the up-right and bottom-left blocks, indicating the regularizer plays a role in sentiment shifting of negators.
- The dots at the up-left and bottom-right respectively indicates the negation effects: changing negative to positive and positive to negative. Typical phrases include *never seems hopelessly* (up-left), *no good scenes* (bottom-right), *not interesting* (bottom-right), etc. There are also some positive/negative phrases shifting to neutral sentiment such as *not so good*, and *not too bad*.
- The dots located at the center indicate that neutral phrases maintain neutral sentiment with negators. Typical phrases include *not at home*, *not here*, where negators typically modify non-sentiment words.

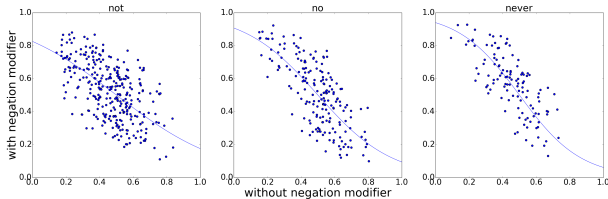


Figure 1: The sentiment shifts with negation words.

³ The score is obtained from the predicted distribution, where 1 means positive and 0 means negative.

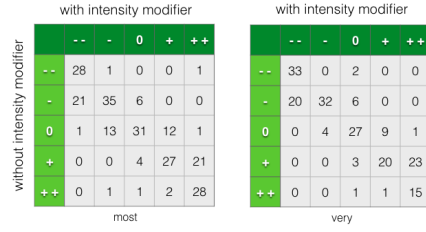


Figure 2: The confusion matrix of sentiment shifting with intensity words.

The Effect of the Intensity Regularizer

To further reveal the linguistic role of intensity words, we perform experiments on the SST dataset, as illustrated in Figure 2. We show the confusion matrix that shows how the sentiment shifts after being modified by intensifiers. The number 20 in the first matrix, for instance, means that there are 20 phrases have a sentiment class of *negative* (-) but shifting to *very negative* (- -) after being modified by an intensity word “very”.

As can be seen from the results, for “most”, there are 21/21/13/12 phrases whose sentiment is shifted from negative to very negative (eg. *most irresponsible picture*), positive to very positive (eg. *most famous author*), neutral to negative (eg. *most plain*), and neutral to positive (eg. *most closely*), respectively. Similar observations can be found with word “very”.

There are also many phrases maintain their sentiment. No surprisingly, for very positive/negative phrases, phrases modified by intensifiers still maintain the strong sentiment. For the left phrases, they fall into three categories: first, words modified by intensifiers are non-sentiment words, such as *most of us*, *most part*; second, intensifiers are not strong enough to shift sentiment, such as *most complex* (from negative to negative), *most traditional* (from *positive* to *positive*); third, our models fail to shift sentiment with intensifiers such as *most vital*, *most resonant film*.

Conclusion and Future Work

We present linguistically regularized LSTMs for sentence-level sentiment classification. The proposed models address the sentiment shifting effect of sentiment, negation, and intensity words to produce linguistically coherent representations. Furthermore, our models are sequence LSTMs which do not depend on a parsing tree-structure and do not require expensive phrase-level annotation to obtain competitive results. Results show that our models are able to address the linguistic role of sentiment, negation, and intensity words.

In order to maintain the simplicity of the proposed models, we do not fully consider the modification scope of negation and intensity words, though we partially address this issue by applying a minimization operator (see Eq. 10, Eq. 13) and bi-directional LSTM. As future work, we plan to apply the linguistic regularizers to tree-LSTM to address the scope issue since the parsing tree is easier to indicate the modification scope explicitly.

References

- [Cho et al. 2014] Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.
- [Chung et al. 2014] Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.
- [Dong et al. 2014] Dong, L.; Wei, F.; Zhou, M.; and Xu, K. 2014. Adaptive multi-compositionality for recursive neural models with applications to sentiment analysis. In *AAAI*. AAAI.
- [Graves, Jaitly, and Mohamed 2013] Graves, A.; Jaitly, N.; and Mohamed, A.-r. 2013. Hybrid speech recognition with deep bidirectional lstm. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, 273–278. IEEE.
- [Hochreiter and Schmidhuber 1997] Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural Computation* 9(8):1735–1780.
- [Hu and Liu 2004] Hu, M., and Liu, B. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 168–177. ACM.
- [Kalchbrenner, Grefenstette, and Blunsom 2014] Kalchbrenner, N.; Grefenstette, E.; and Blunsom, P. 2014. A convolutional neural network for modelling sentences. In *ACL*, 655–665.
- [Kennedy and Inkpen 2006] Kennedy, A., and Inkpen, D. 2006. Sentiment classification of movie reviews using contextual valence shifters. *Computational intelligence* 22(2):110–125.
- [Kim 2014] Kim, Y. 2014. Convolutional neural networks for sentence classification. In *EMNLP*, 1746–1751.
- [Kiritchenko and Mohammad 2016] Kiritchenko, S., and Mohammad, S. M. 2016. Sentiment composition of words with opposing polarities. In *NAACL*.
- [Malandrakis et al. 2013] Malandrakis, N.; Potamianos, A.; Iosif, E.; and Narayanan, S. 2013. Distributional semantic models for affective text analysis. *IEEE Transactions on Audio, Speech, and Language Processing* 21(11):2379–2392.
- [Mikolov 2012] Mikolov, T. 2012. Statistical language models based on neural networks. *Presentation at Google, Mountain View, 2nd April*.
- [Pang and Lee 2005] Pang, B., and Lee, L. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL*, 115–124.
- [Pang and Lee 2008] Pang, B., and Lee, L. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval* 2(1-2):1–135.
- [Pang, Lee, and Vaithyanathan 2002] Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *ACL*, 79–86.
- [Polanyi and Zaenen 2006] Polanyi, L., and Zaenen, A. 2006. Contextual valence shifters. In *Computing attitude and affect in text: Theory and applications*. Springer. 1–10.
- [Qian et al. 2015] Qian, Q.; Tian, B.; Huang, M.; Liu, Y.; Zhu, X.; and Zhu, X. 2015. Learning tag embeddings and tag-specific composition functions in recursive neural network. In *ACL*, volume 1, 1365–1374.
- [Socher et al. 2011] Socher, R.; Pennington, J.; Huang, E. H.; Ng, A. Y.; and Manning, C. D. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *EMNLP*, 151–161.
- [Socher et al. 2013] Socher, R.; Perelygin, A.; Wu, J. Y.; Chuang, J.; Manning, C. D.; Ng, A. Y.; and Potts, C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, 1631–1642.
- [Taboada et al. 2011] Taboada, M.; Brooke, J.; Tofiloski, M.; Voll, K.; and Stede, M. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics* 37(2):267–307.
- [Tai, Socher, and Manning 2015] Tai, K. S.; Socher, R.; and Manning, C. D. 2015. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*.
- [Turney 2002] Turney, P. D. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *ACL*, 417–424.
- [Wang, Zhang, and Lan 2016] Wang, F.; Zhang, Z.; and Lan, M. 2016. Ecnv at semeval-2016 task 7: An enhanced supervised learning method for lexicon sentiment intensity ranking. *Proceedings of SemEval* 491–496.
- [Wei, Wu, and Lin 2011] Wei, W.-L.; Wu, C.-H.; and Lin, J.-C. 2011. A regression approach to affective rating of chinese words from anew. In *Affective Computing and Intelligent Interaction*. Springer. 121–131.
- [Wilson, Wiebe, and Hoffmann 2005] Wilson, T.; Wiebe, J.; and Hoffmann, P. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *EMNLP*, 347–354.
- [Zhu et al. 2014] Zhu, X.; Guo, H.; Mohammad, S.; and Kiritchenko, S. 2014. An empirical study on the effect of negation words on sentiment. In *ACL*, 304–313.
- [Zhu, Sobhani, and Guo 2015] Zhu, X.; Sobhani, P.; and Guo, H. 2015. Long short-term memory over recursive structures. In *ICML*, 1604–1612.